Protein Folding

Unfolded

⇕ Rapid (< 1s)

Molten Globule State

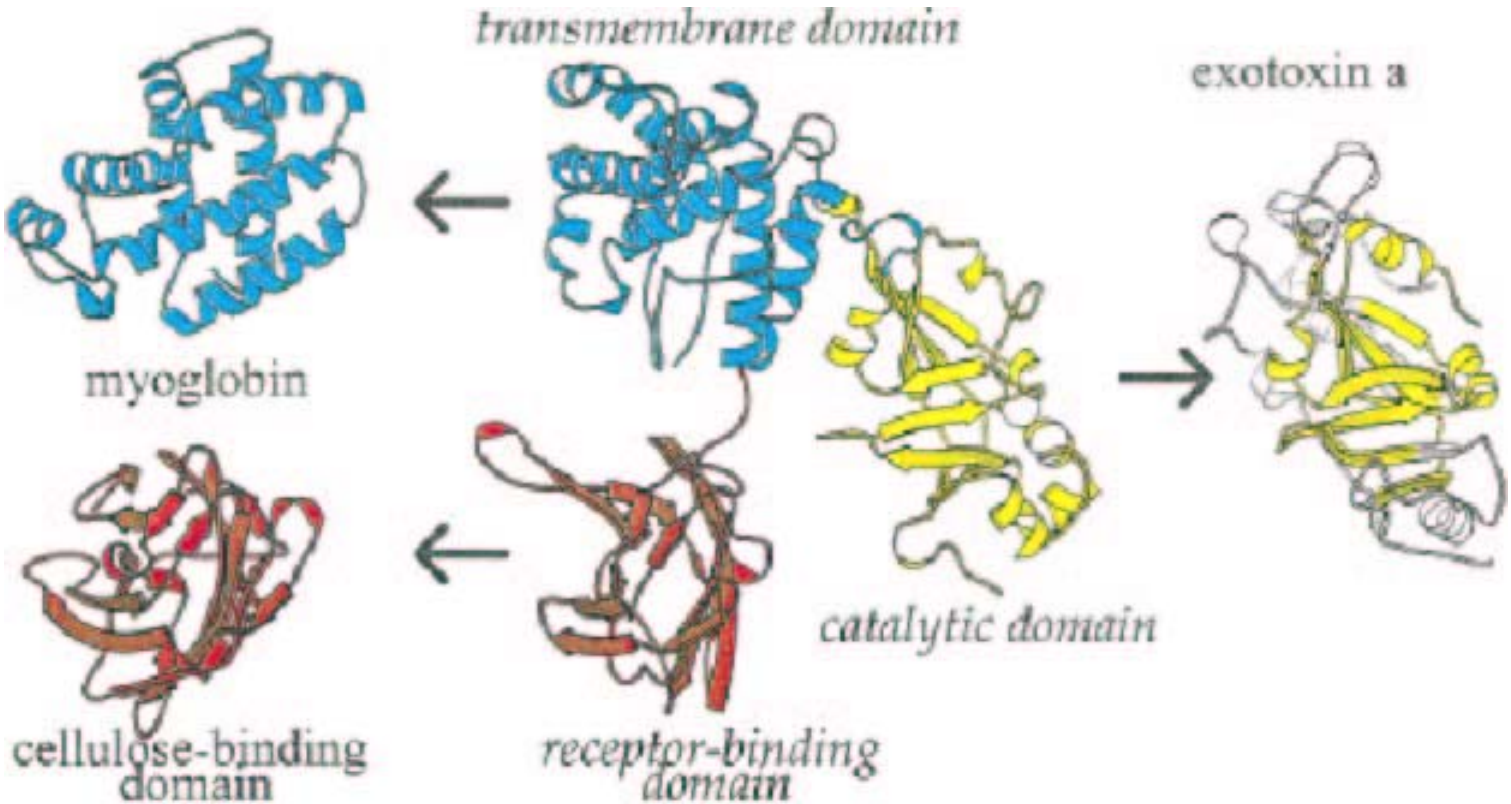
⇕ Slow (1 – 1000 s)

Folded Native State

- How to find minimum energy configuration?

Modular Nature of Protein Structures

Example: Diphtheria Toxin



Structural Classification of Proteins

- SCOP (Structural Classification of Proteins)
  - Based on structurla & evolutionary relationships.
  - Contains ~ 40,000 domains
  - Classes (groups of folds), Folds (proteins sharing folds), Families (proteins related by function/evolution), Superfamilies (distantly related proteins)

SCOP Family View

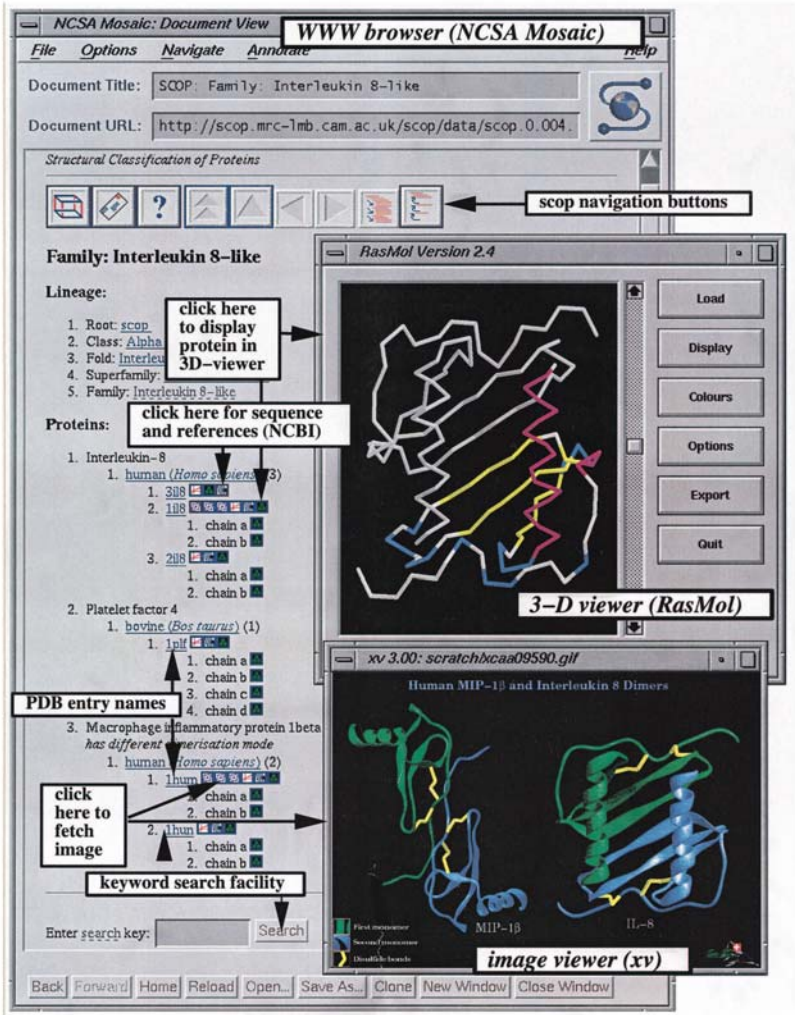


Figure 2. A typical scop session is shown on a unix workstation. A scop page, of the Interleukin 8-like family, is displayed by the *WWW browser program (NCSA Mosaic)* (Schatz & Hardin, 1994). Navigating through the tree structure is accomplished by selecting any underlined entry, by clicking on buttons (at the top of each page) and by keyword searching (at the bottom of each page). The static image comparing two proteins in this family was downloaded by clicking on the icon indicated and is displayed by image-viewer program *xv*. By clicking on one of the green icons, commands were sent to a molecular viewer program (*RasMol*) written by Roger Sayle (Sayle, 1994), instructing it to automatically display the relevant PDB file and colour the domain in question by secondary structure. Since sending large PDB files over the network can be slow, this feature of scop can be configured to use local copies of PDB files if they are available. Equivalent WWW browsers, image-display programs and molecular viewers are also available free for Windows-PC and Macintosh platforms.

CATH: Protein Structure Classification

- Semi-automatic classification; ~36K domains
- 4 levels of classification:
  - Class (C), depends on sec. Str. Content
  - Architecture (A), orientation of sec. Str.
  - Topolgy (T), topological connections &
  - Homologous Superfamily (H), similar str and functions.

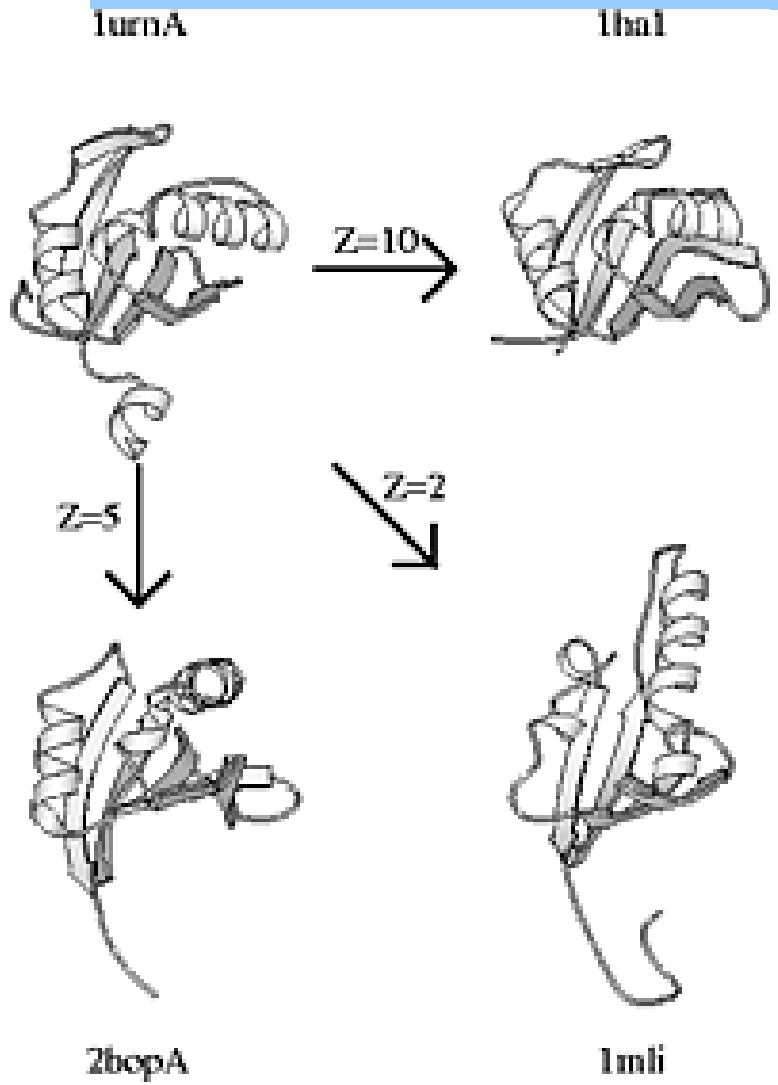
DALI Domain Dictionary

- Completely automated; 3724 domains
- Criteria of compactness & recurrence
- Each domain is assigned a Domain Classification number DC_l_m_n_p representing fold space attractor region (l), globular folding topology (m), functional family (n) and sequence family (p).

5 Fold Space classes

Attractor 1 can be characterized as alpha/beta,
attractor 2 as all-beta, attractor 3 as all-alpha,
attractor 5 as alpha-beta meander (1mli), and
attractor 4 contains antiparallel beta-barrels e.g. OB-fold (1prtF).

Fold Types & Neighbors



Structural neighbours of 1urnA (top left). 1mli (bottom right) has the same topology even though there are shifts in the relative orientation of secondary structure elements.

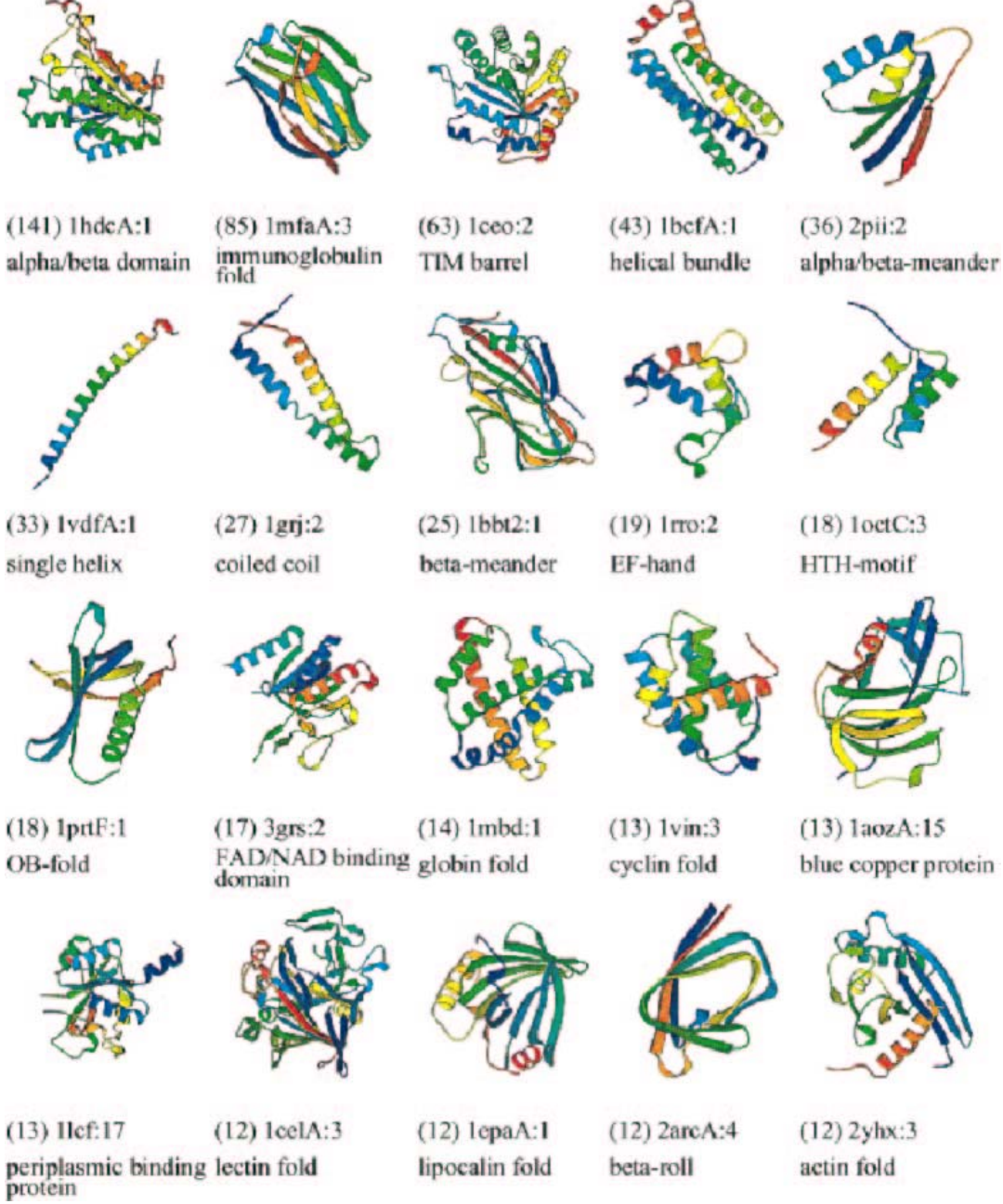
Sequence Alignment of Fold Neighbors

B

```
1urnA --RPNHTIYINNLNEKI----KKDELKKSLHAIFSRFG---QILDILV-SRS---LKM---
Z=10          *      *                 *   *         *  *          *
1ha1  ahLTVKKIFVGGIKEDT--------EEHHLRDYFEQYG---KIEVIEI-MTDrgsGKK---
Z=5            *
2bopA ----sCFALIS-GTANQ-----vKCYRFRVKKNHRHR-----YENCTTtWFT---Vadnga
Z=2                                       *
1mli  ---mlFHVKMTVKLpvdmdpakatqlkadeKELAQRlgregTWRHLWR-IAG---------

1urnA ----RGQAFVIFKEV--SSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKM---------
Z=10      ** *** *       *                     *
1ha1  ----RGFAFVTFDDH--DSVDKIVIQ-kYHTVNGHNCEVRKAL------------------
Z=5          *     *      *      *       *     * *
2bopA ergqQAQILITFGSP--SQRQDFLKHVPLPP----GMNISGF-----tASLDf--------
Z=2                *             *      **      * *
1mli  ----HYANYSVFDVpsvEALHDTLMQLpLFPY----MDIEVD-----gLCRHpssihsddr
```

Frequent Fold Types



Gene Expression

- Process of transcription and/or translation of a gene is called gene expression.
- Every cell of an organism has the same genetic material, but different genes are expressed at different times.
- Patterns of gene expression in a cell is indicative of its state.

Hybridization

- If two complementary strands of DNA or mRNA are brought together, under appropriate experimental conditions they will hybridize.
- A hybridizes to B $\Rightarrow$
  - A is reverse complementary to B, or
  - A is reverse complementary to a subsequence of B.
- It is possible to experimentally verify whether A hybridizes to B, by labeling A or B with a radioactive or fluorescent tag, followed by excitation by laser.

Measuring gene expression

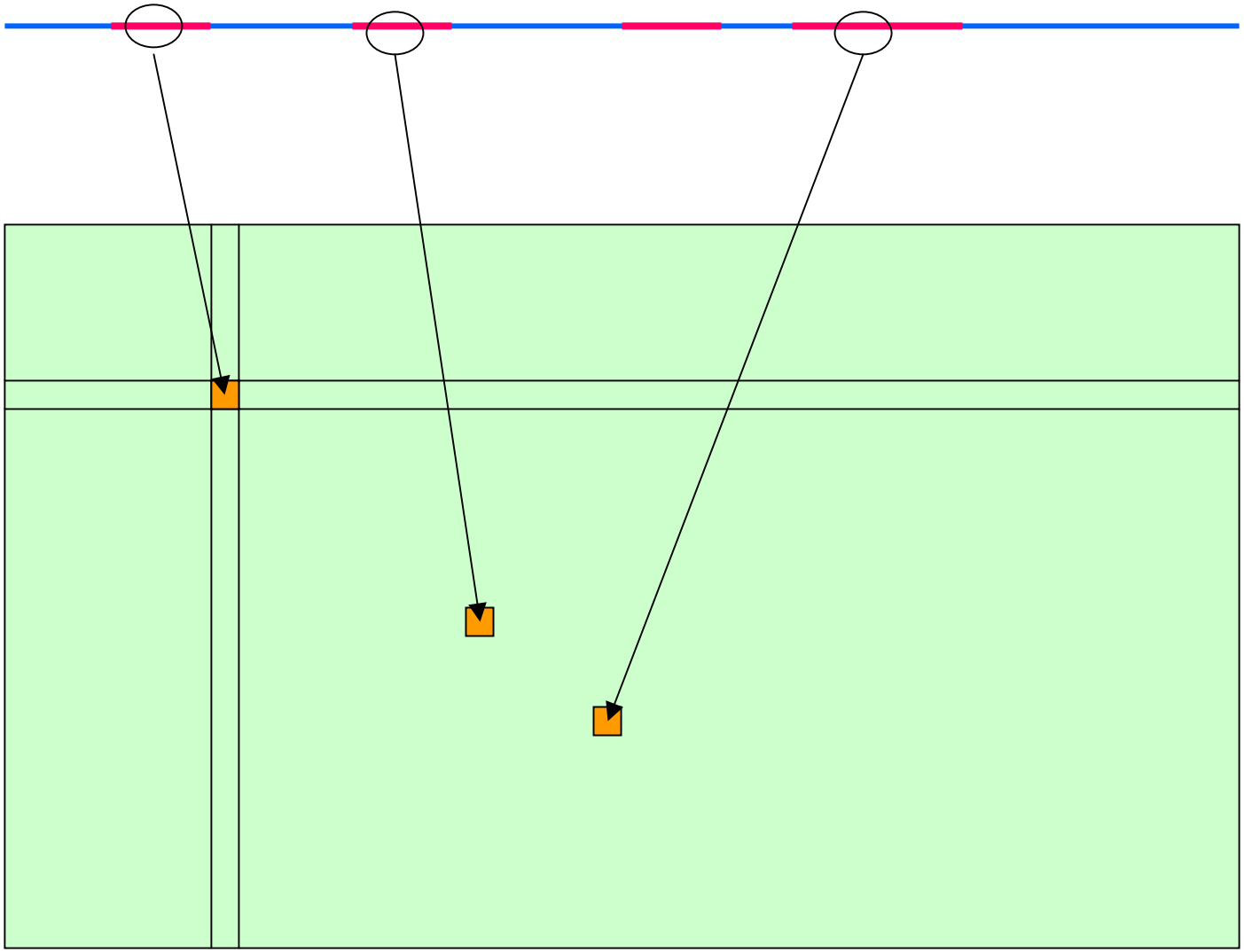
- Gene expression for a single gene can be measured by extracting mRNA from the cell and doing a simple hybridization experiment.
- Given a sample of cells, gene expression for every gene can be measured using a single microarray experiment.

Microarray/DNA chip technology

- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between tissues, between drugs, across disease states

Microarray Data

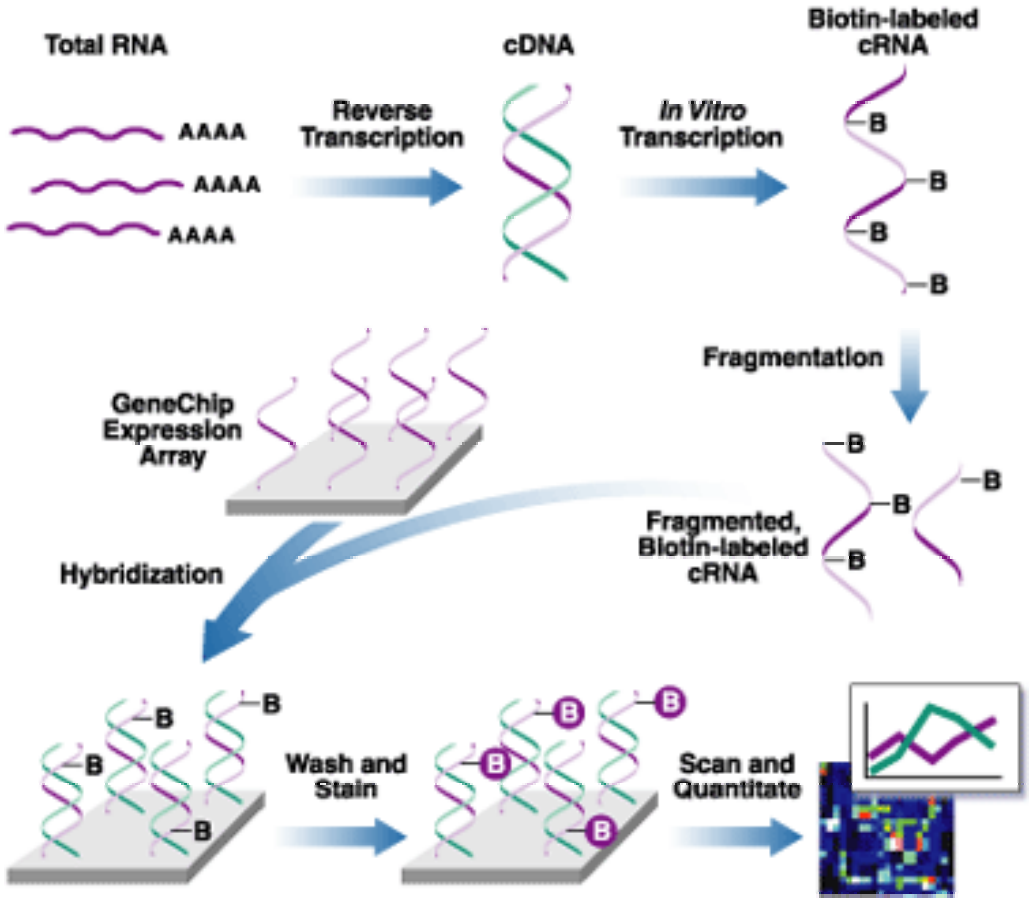
Gene	Expression Level
Gene1	
Gene2	
Gene3	
…	



Microarray/DNA chips (Simplified)

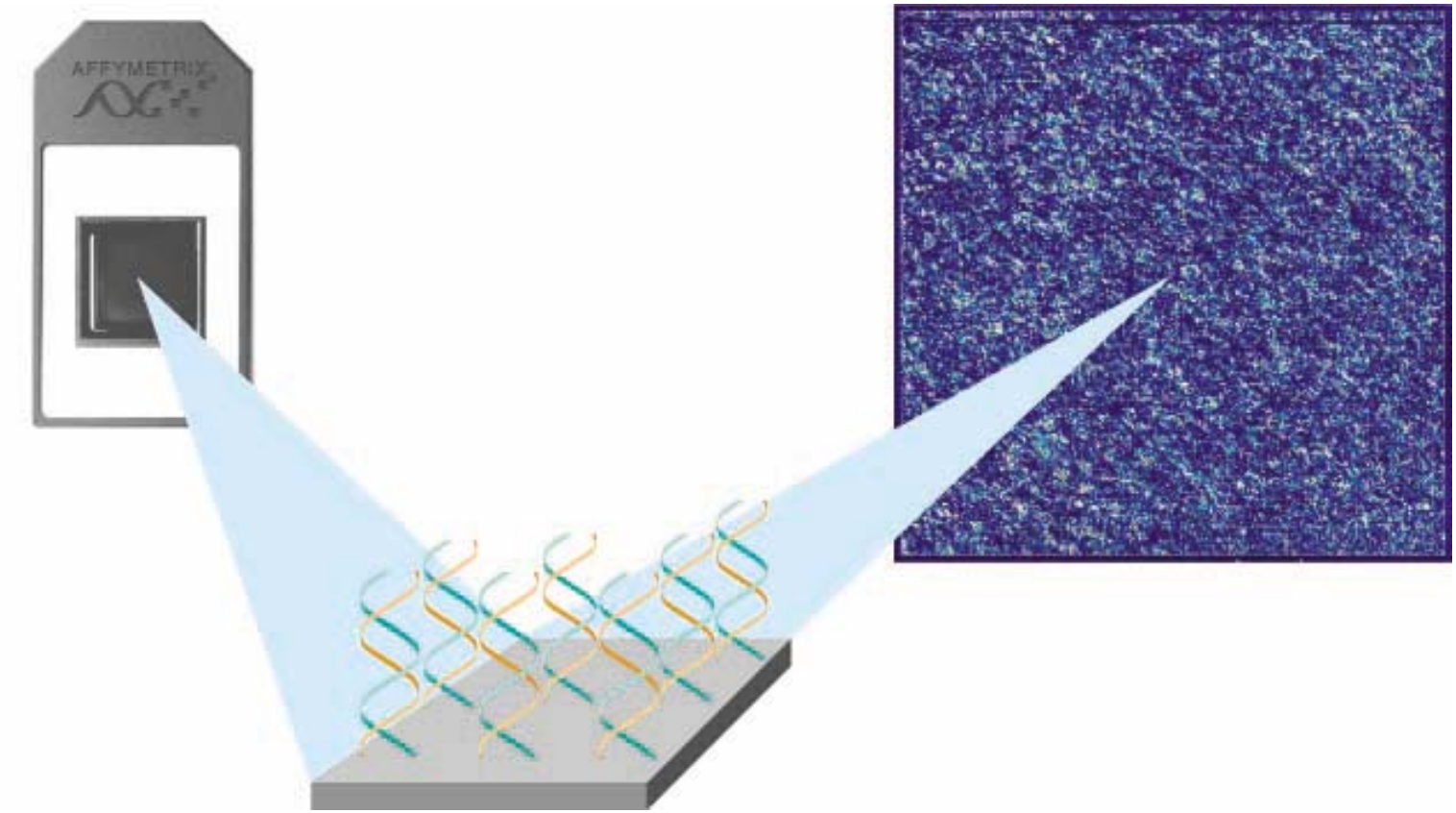
- Construct probes corresponding to reverse complements of genes of interest.
- Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- Extract mRNA from sample cells and label them.
- Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.
- Wash off unhybridized material.
- Use optical detector to measure amount of fluorescence from each spot.

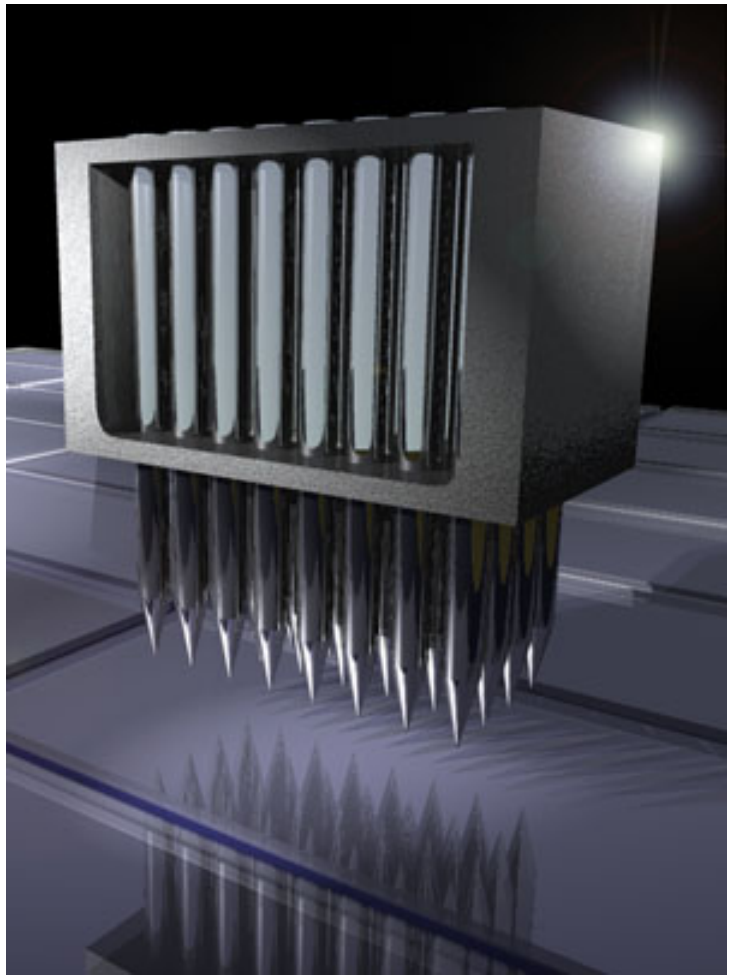
Affymetrix DNA chip schematic



www.affymetrix.com

DNA Chips & Images





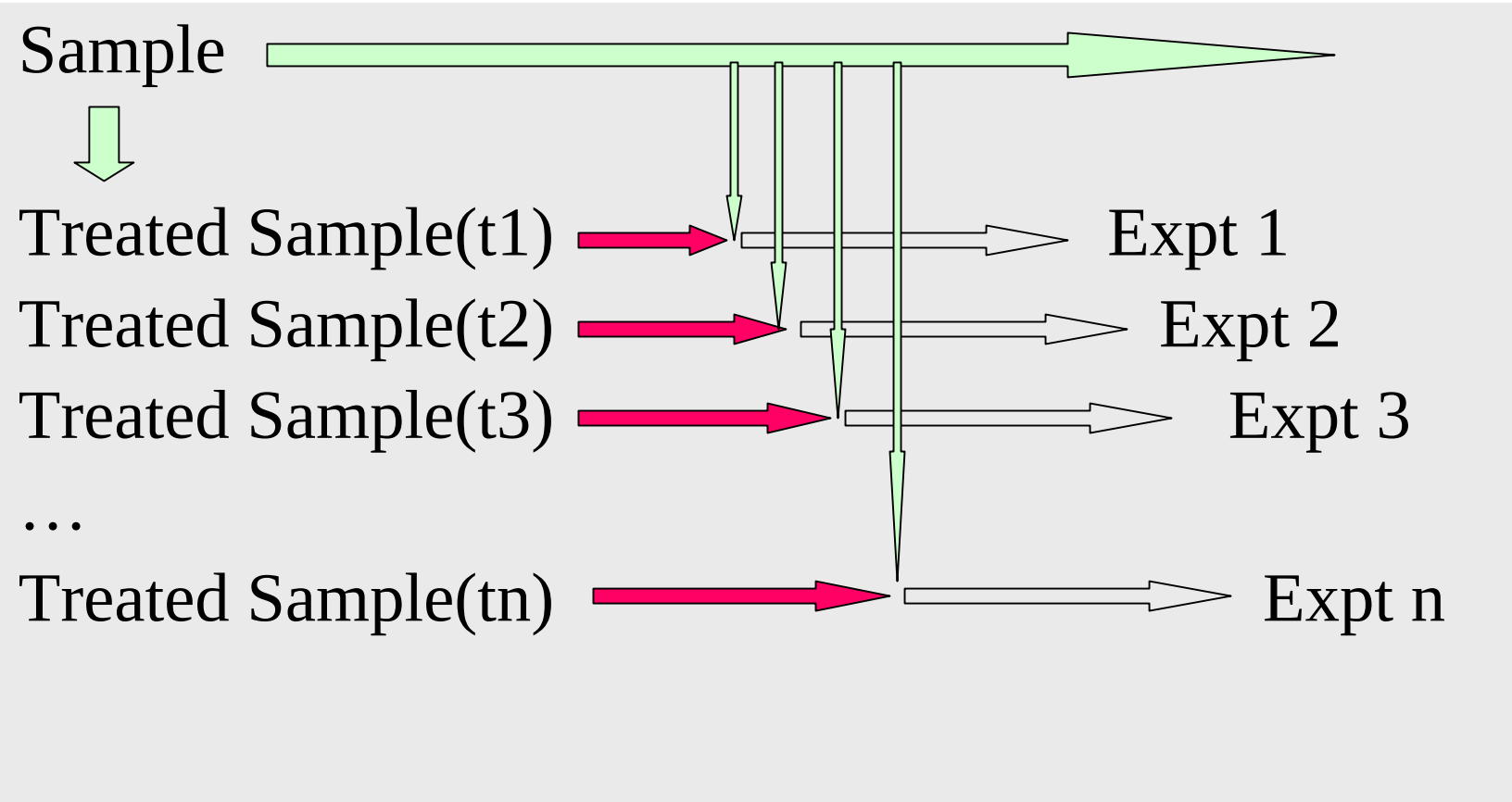
Microarrays: competing technologies

- Affymetrix & Synteni/Stanford
- Differ in:
  - method to place DNA: Spotting vs. photolithography
  - Length of probe
  - Complete sequence vs. series of fragments

How to compare 2 cell samples?

- mRNA from sample 1 is extracted and labeled with a red fluorescent dye.
- mRNA from sample 2 is extracted and labeled with a green fluorescent dye.
- Mix the samples and apply it to every spot on the microarray. Hybridize sample mixture to probes.
- Use optical detector to measure the amount of green and red fluorescence at each spot.

Studying effect of a treatment over time



Sources of Variations & Errors

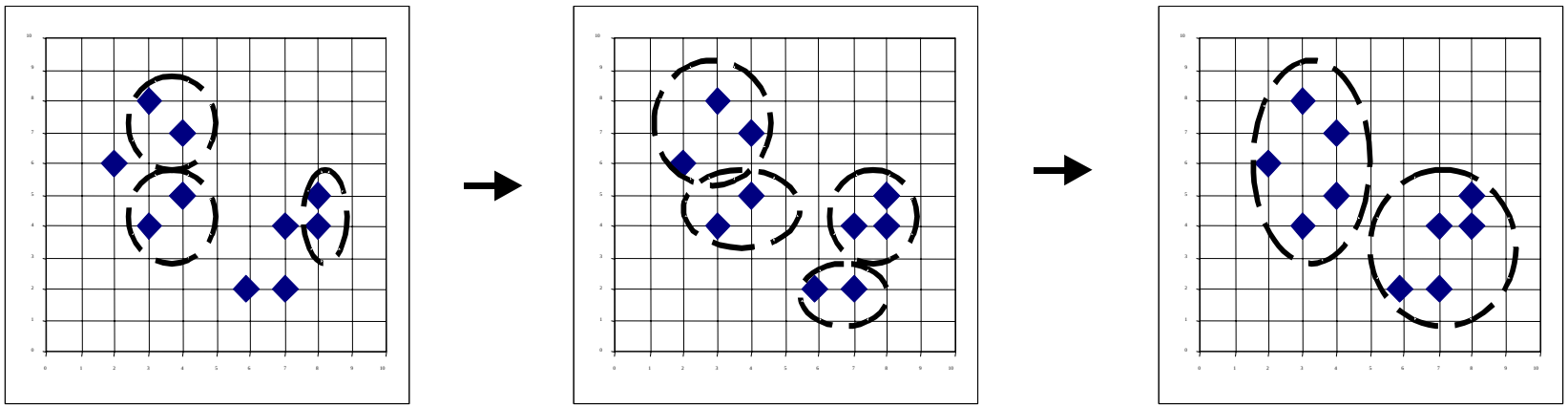
- Variations in cells/individuals.
- Variations in mRNA extraction, isolation, introduction of dye, variation in dye incorporation, dye interference.
- Variations in probe concentration, probe amounts, substrate surface characteristics
- Variations in hybridization conditions and kinetics
- Variations in optical measurements, spot misalignments, discretization effects, noise due to scanner lens and laser irregularities
- Cross-hybridization of sequences with high sequence identity.
- Limit of factor 2 in precision of results.

Need to Normalize data

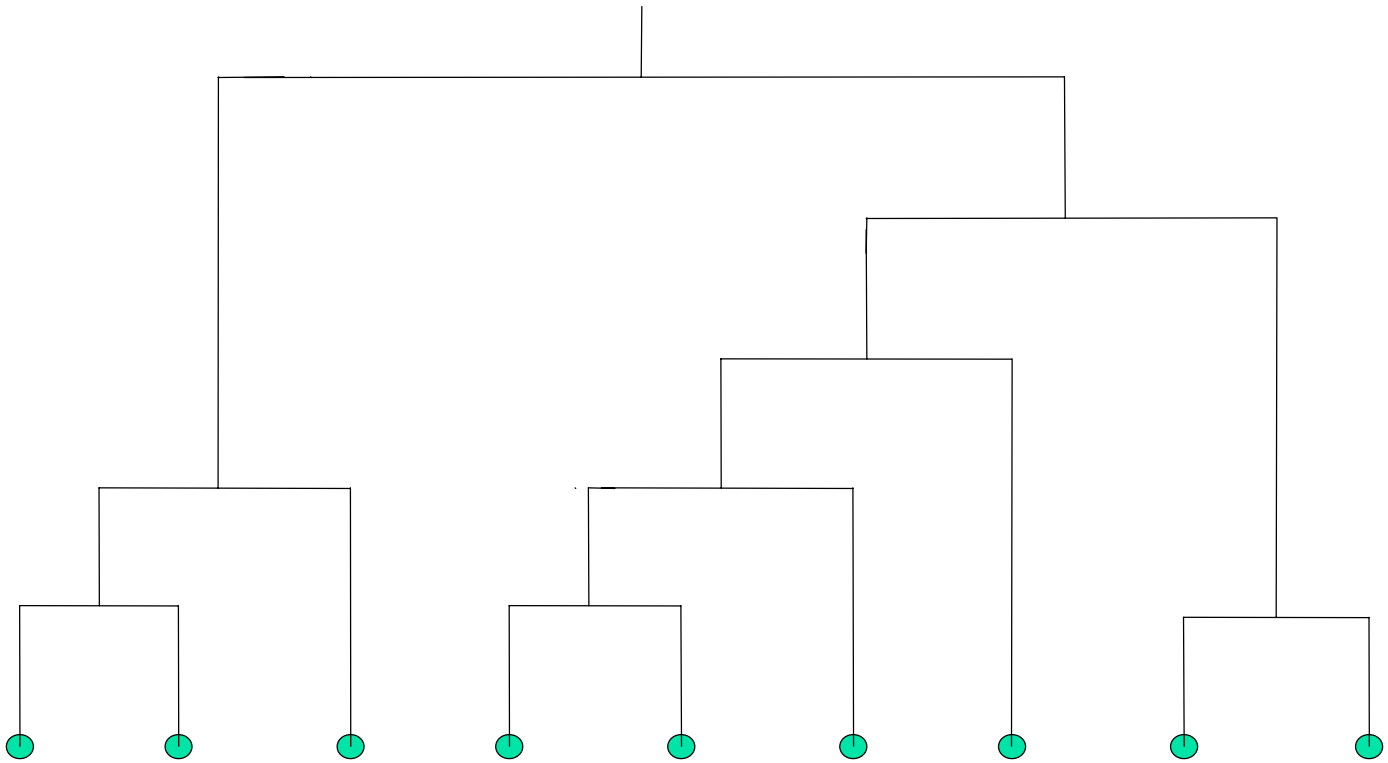
Clustering

- Clustering is a general method to study patterns in gene expressions.
- Several known methods:
  - Hierarchical Clustering (Bottom-Up Approach)
  - K-means Clustering (Top-Down Approach)
  - Self-Organizing Maps (SOM)

Hierarchical Clustering: Example



A Dendrogram



Hierarchical Clustering [Johnson, SC, 1967]

- Given n points in R^d, compute the distance between every pair of points
- While (not done)
  - Pick closest pair of points s_i and s_j and make them part of the same cluster.
  - Replace the pair by an average of the two s_{ij}

Try the applet at:
http://www.cs.mcgill.ca/~papou/#applet

Distance Metrics

- For clustering, define a distance function:
  - Euclidean distance metrics

$$D_k(X,Y) = \left[\sum_{i=1}^{d} (X_i - Y_i)^k \right]^{1/k}$$

k=2: Euclidean Distance

  - Pearson correlation coefficient

$$\rho_{xy} = \frac{1}{d} \sum_{i=1}^{d} \left(\frac{X_i - \overline{X}}{\sigma_x} \right) \left(\frac{Y_i - \overline{Y}}{\sigma_y} \right)$$

$-1 \leq \rho_{xy} \geq 1$

EXHIBIT 3.4 *Joint Probability Model for the Ratings of Two People*

(a) $\rho_{XY} = 0$

	y			
x	1	2	3	Total
3	1/9	1/9	1/9	1/3
2	1/9	1/9	1/9	1/3
1	1/9	1/9	1/9	1/3
Total	1/3	1/3	1/3	1

(b) $\rho_{XY} = \frac{1}{2}$

	y			
x	1	2	3	Total
3	1/18	1/18	4/18	1/3
2	1/18	4/18	1/18	1/3
1	4/18	1/18	1/18	1/3
Total	1/3	1/3	1/3	1

(c) $\rho_{XY} = -\frac{1}{2}$

	y			
x	1	2	3	Total
3	4/18	1/18	1/18	1/3
2	1/18	4/18	1/18	1/3
1	1/18	1/18	4/18	1/3
Total	1/3	1/3	1/3	1

(d) $\rho_{XY} = \frac{4}{9}$

	y			
x	1	2	3	Total
3	1/27	2/27	6/27	1/3
2	2/27	5/27	2/27	1/3
1	6/27	2/27	1/27	1/3
Total	1/3	1/3	1/3	1

(e) $\rho_{XY} = -\frac{5}{9}$

	y			
x	1	2	3	Total
3	6/27	2/27	1/27	1/3
2	2/27	5/27	2/27	1/3
1	1/27	2/27	6/27	1/3
Total	1/3	1/3	1/3	1

(f) $\rho_{XY} = \frac{2}{3}$

	y			
x	1	2	3	Total
3	1/36	2/36	9/36	1/3
2	2/36	8/36	2/36	1/3
1	9/36	2/36	1/36	1/3
Total	1/3	1/3	1/3	1

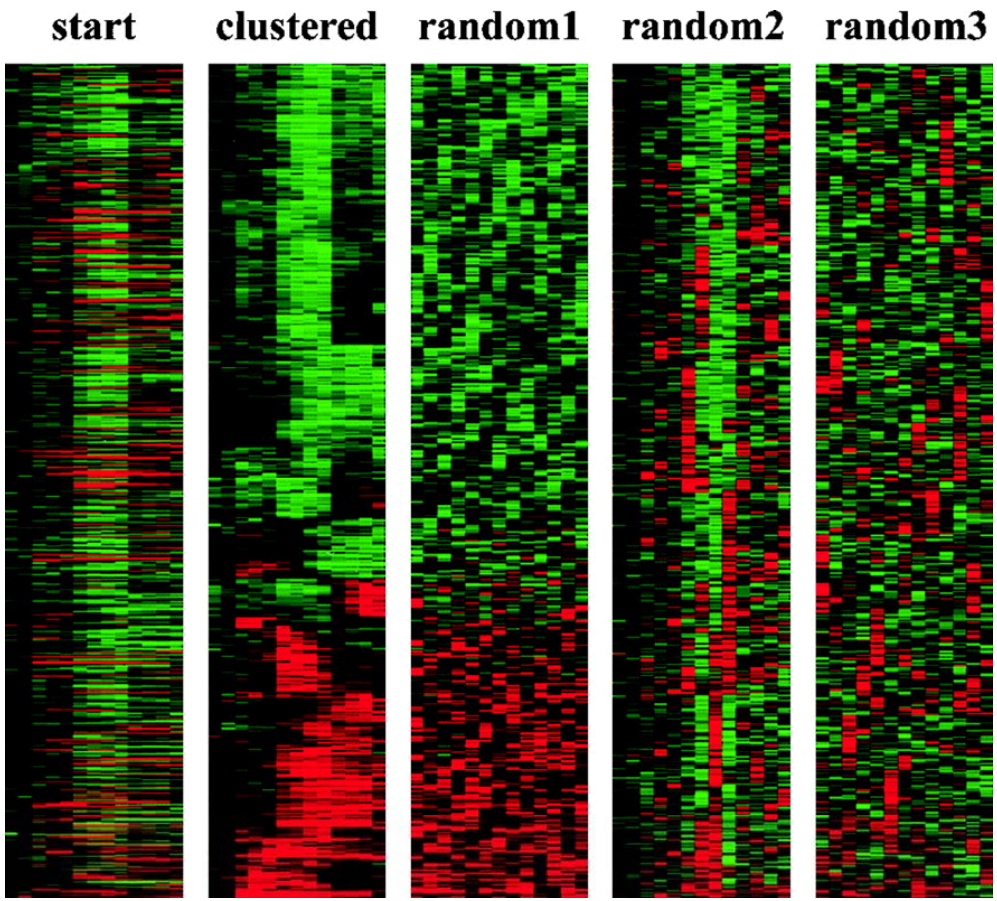
(g) $\rho_{XY} = -\frac{1}{3}$

	y			
x	1	2	3	Total
3	9/36	2/36	1/36	1/3
2	2/36	8/18	2/18	1/3
1	1/36	2/36	9/36	1/3
Total	1/3	1/3	1/3	1

Clustering of gene expressions

- Represent each gene as a vector or a point in d-space where d is the number of arrays or experiments being analyzed.

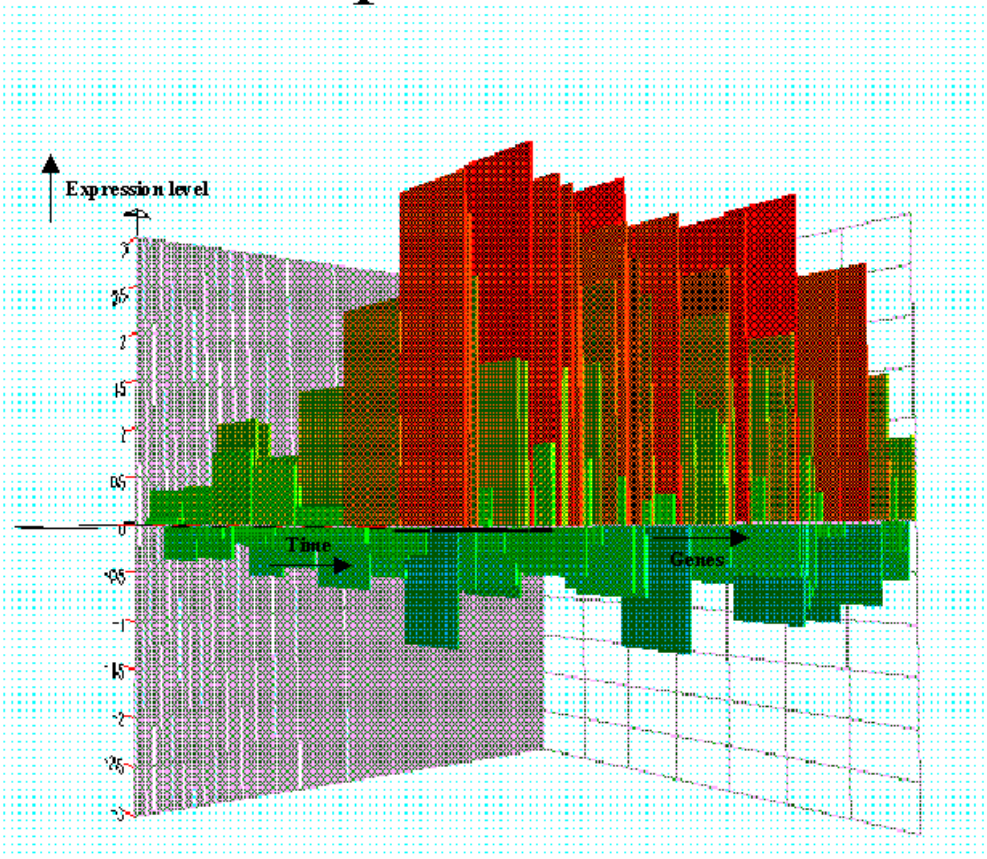
Clustering Random vs. Biological Data



From *Eisen MB, et al, PNAS 1998 95(25):14863-8*

European Bioinformatics Institute

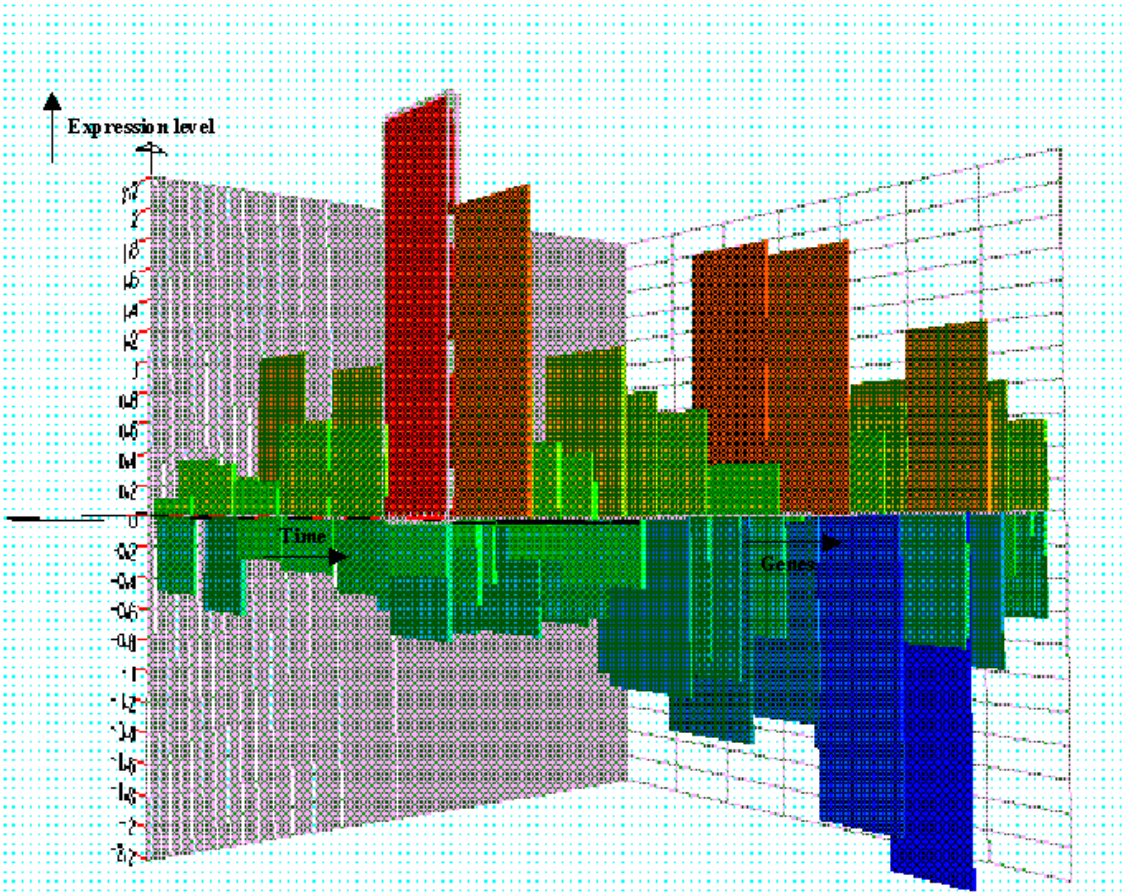
Expression Profiles for Respiration Genes



Genes show general increase in expression level during diauxic shift

European Bioinformatics Institute

Expression Profiles for Fermentation Genes



Bar two exceptions, genes show general decrease in expression level during diauxic shift

European Bioinformatics Institute

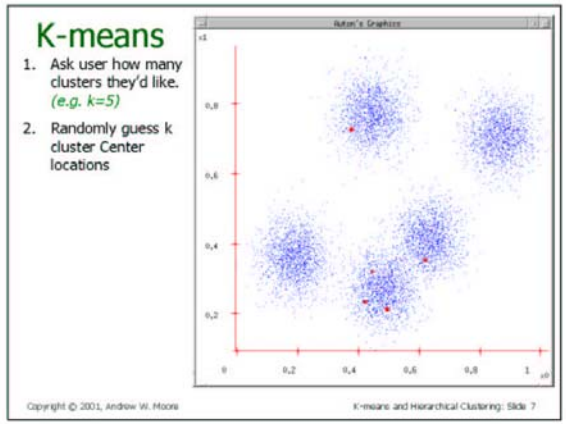
Observations

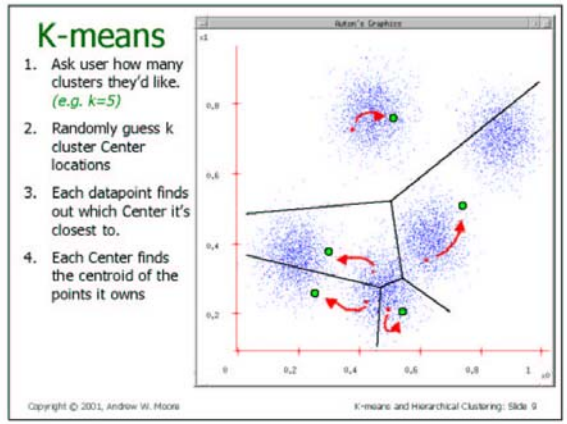
- As glucose was depleted - Marked change in the global pattern of gene expression
- ~50% of differentially expressed genes have unknown function
- Genes with similar expression profiles had common promoters
- Expression patterns observed match those observed in other types of experiments

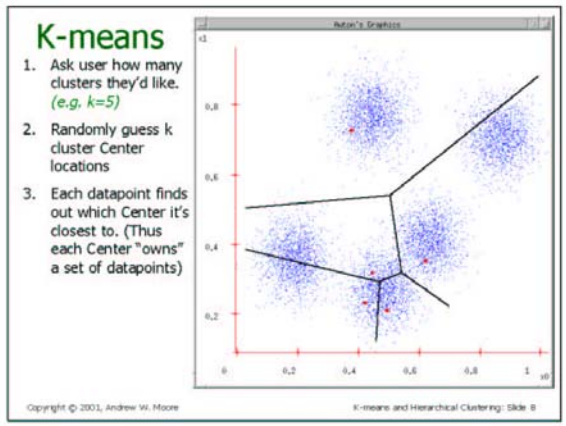
K-Means Clustering: Example

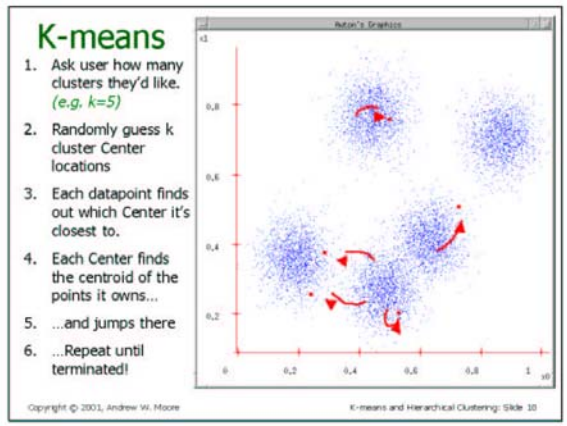
Example from Andrew Moore's tutorial on Clustering.

Start


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 7


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 9


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 8


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 10

4

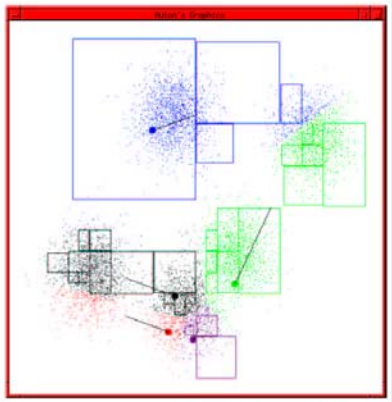
5

K-means Start

Advance apologies: in Black and White this example will deteriorate

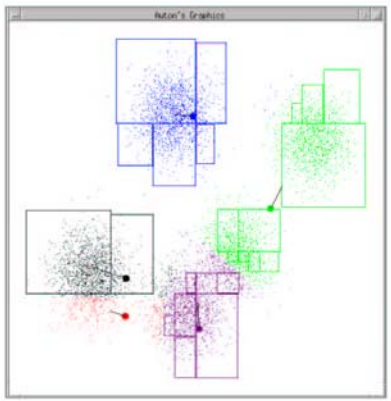
Example generated by Dan Pelleg's super-duper fast K-means system:

Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999, (KDD99) (available on www.autonlab.org/pap.html)



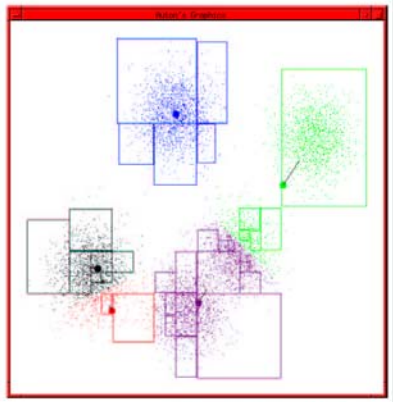
K-means and Hierarchical Clustering: Slide 11

K-means continues …



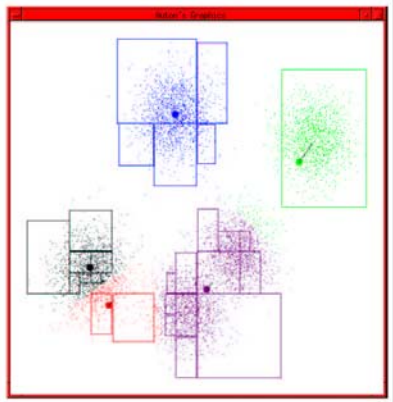
K-means and Hierarchical Clustering: Slide 12

K-means continues …



K-means and Hierarchical Clustering: Slide 13

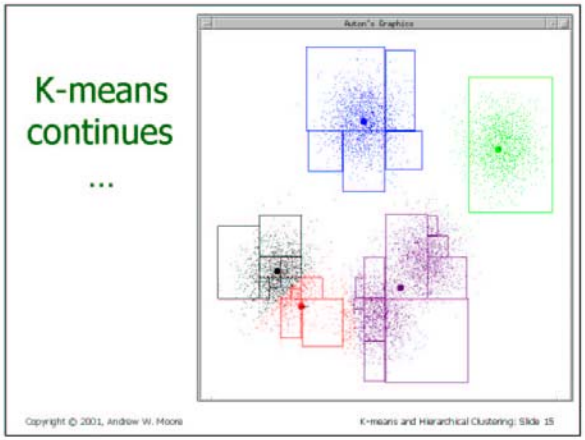
K-means continues …

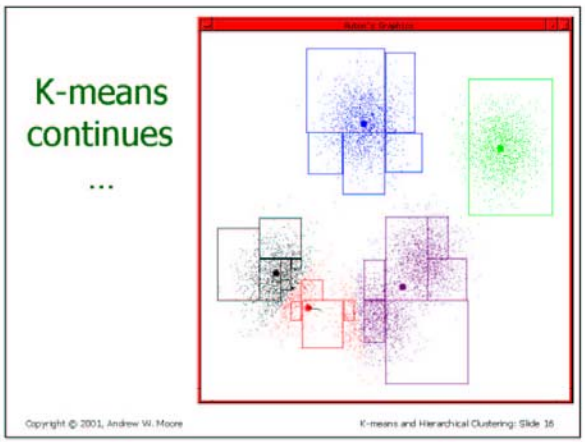


K-means and Hierarchical Clustering: Slide 14

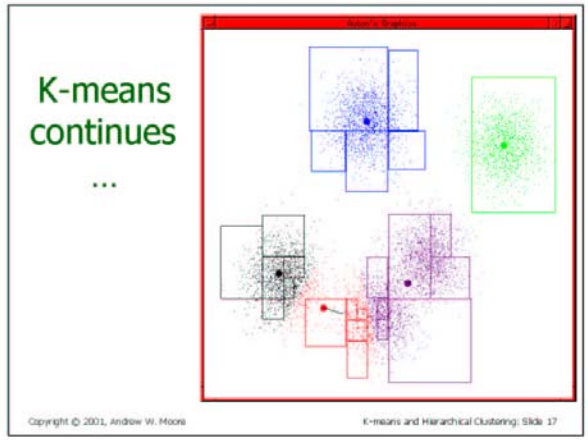
6

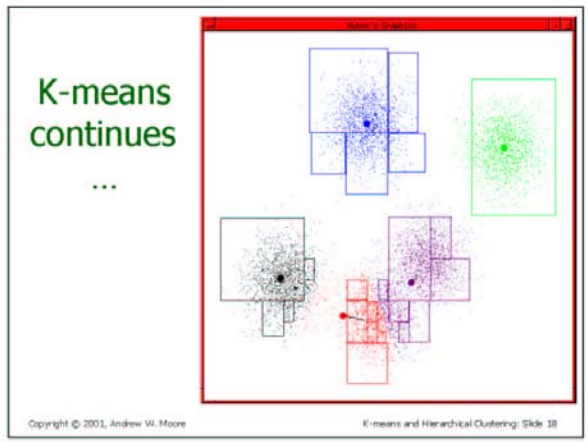
7


K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 15


K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 16

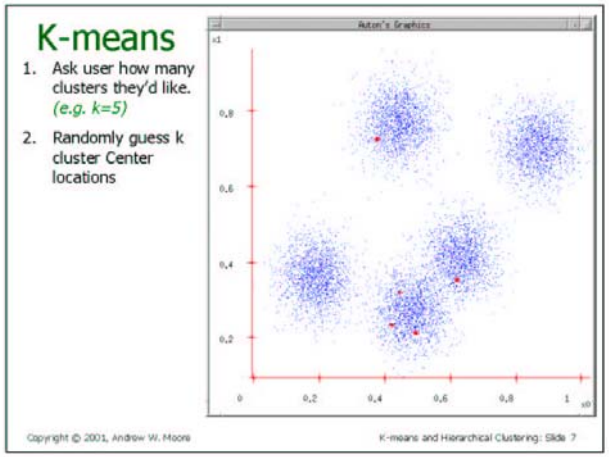
8

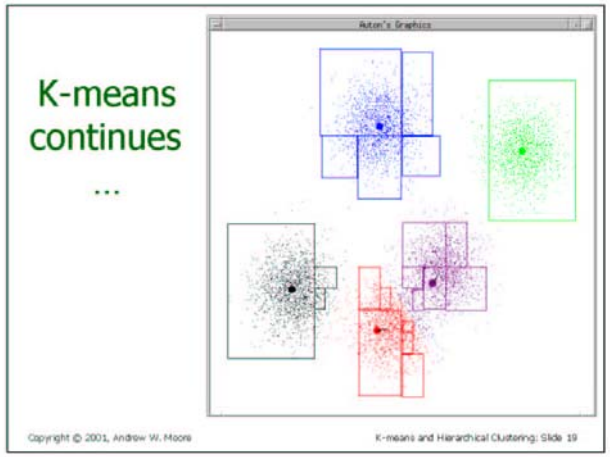

K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 17

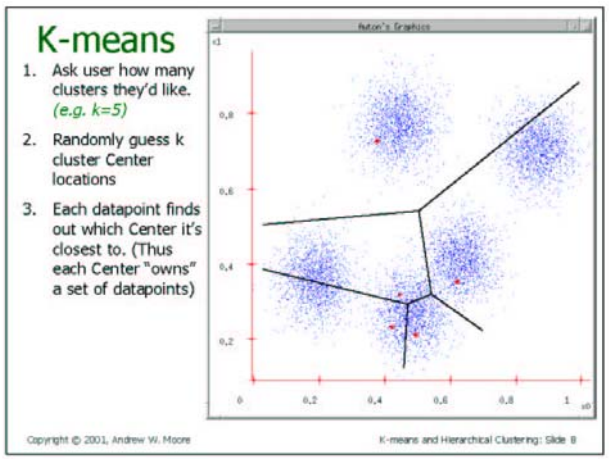

K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 18

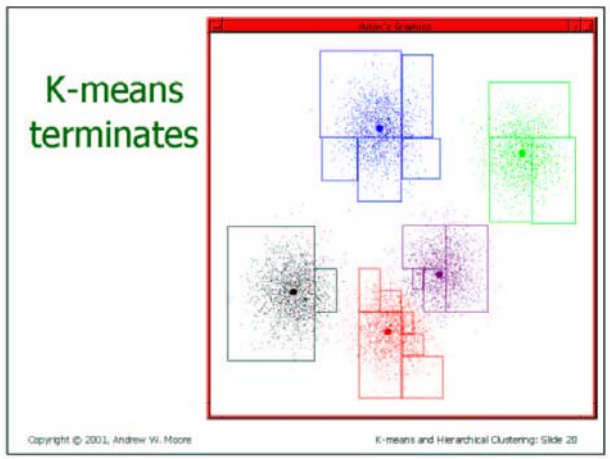
9

Start









End

4

10

K-Means Clustering [McQueen ’67]

Repeat

- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://www.cs.mcgill.ca/~bonnef/project.html

Comparisons

- Hierarchical clustering
  - Number of clusters not preset.
  - Complete hierarchy of clusters
  - Not very robust, not very efficient.
- K-Means
  - Need definition of a mean. Categorical data?
  - More efficient and often finds optimum clustering.

Functionally related genes behave similarly across experiments

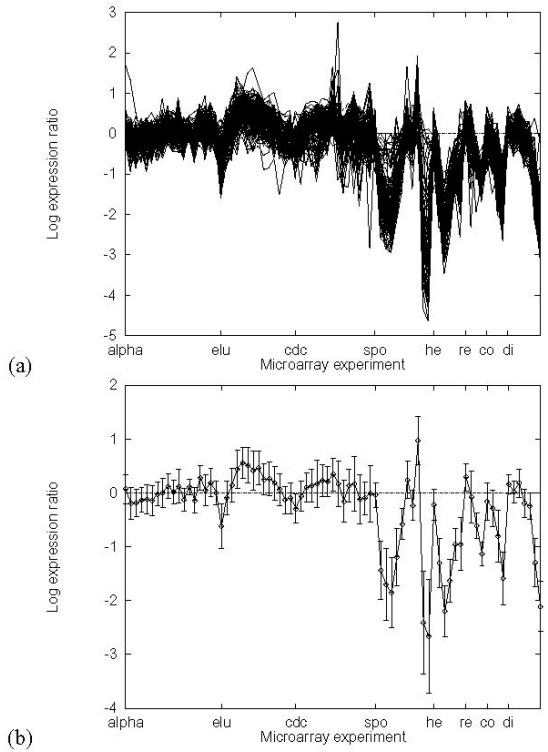


Figure 1: **Expression profiles of the cytoplasmic ribosomal proteins.** Figure (a) shows the expression profiles from the data in [Eisen et al., 1998] of 121 cytoplasmic ribosomal proteins, as classified by MYGD [MYGD, 1999]. The logarithm of the expression ratio is plotted as a function of DNA microarray experiment. Ticks along the X-axis represent the beginnings of experimental series. They are, from left to right, cell division cycle after synchronization with α factor arrest (alpha), cell division cycle after synchronization by centrifugal elutriation (elu), cell division cycle measured using a temperature sensitive *cdc15* mutant (cdc), sporulation (spo), heat shock (he), reducing shock (re), cold shock (co), and diauxic shift (di). Sporulation is the generation of a yeast spore by meiosis. Diauxic shift is the shift from anaerobic (fermentation) to aerobic (respiration) metabolism. The medium starts rich in glucose, and yeast cells ferment, producing ethanol. When the glucose is used up, they switch to ethanol as a source for carbon. Heat, cold, and reducing shock are various ways to stress the yeast cell. Figure (b) shows the average, plus or minus one standard deviation, of the data in Figure (a).

Self-Organizing Maps [Kohonen]

- Kind of neural network.
- Clusters data and find complex relationships between clusters.
- Helps reduce the dimensionality of the data.
- Map of 1 or 2 dimensions produced.
- Unsupervised Clustering
- Like K-Means, except for visualization

SOM Architectures

- 2-D Grid
- 3-D Grid
- Hexagonal Grid

SOM Algorithm

- Select SOM architecture, and initialize weight vectors and other parameters.
- **While** (stopping condition not satisfied) **do** for each input point x
  - winning node q has weight vector closest to x.
  - Update weight vector of q and its neighbors.
  - Reduce neighborhood size and learning rate.

SOM Algorithm Details

- Distance between x and weight vector: $\|x - w_i\|$
- Winning node: $q(x) = \min_i \|x - w_i\|$
- Weight update function (for neighbors):

$$w_i(k+1) = w_i(k) + \mu(k, x, i)[x(k) - w_i(k)]$$

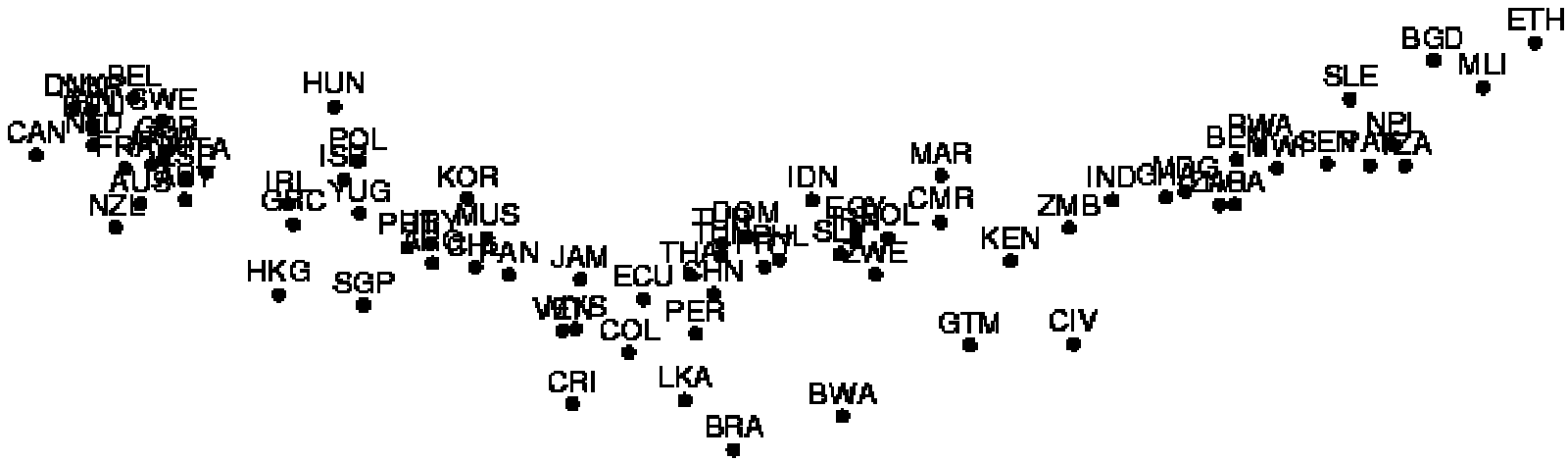
- Learning rate:

$$\mu(k, x, i) = \eta_0(k) \exp\left(\frac{-\|r_i - r_{q(x)}\|^2}{\sigma^2} \right)$$

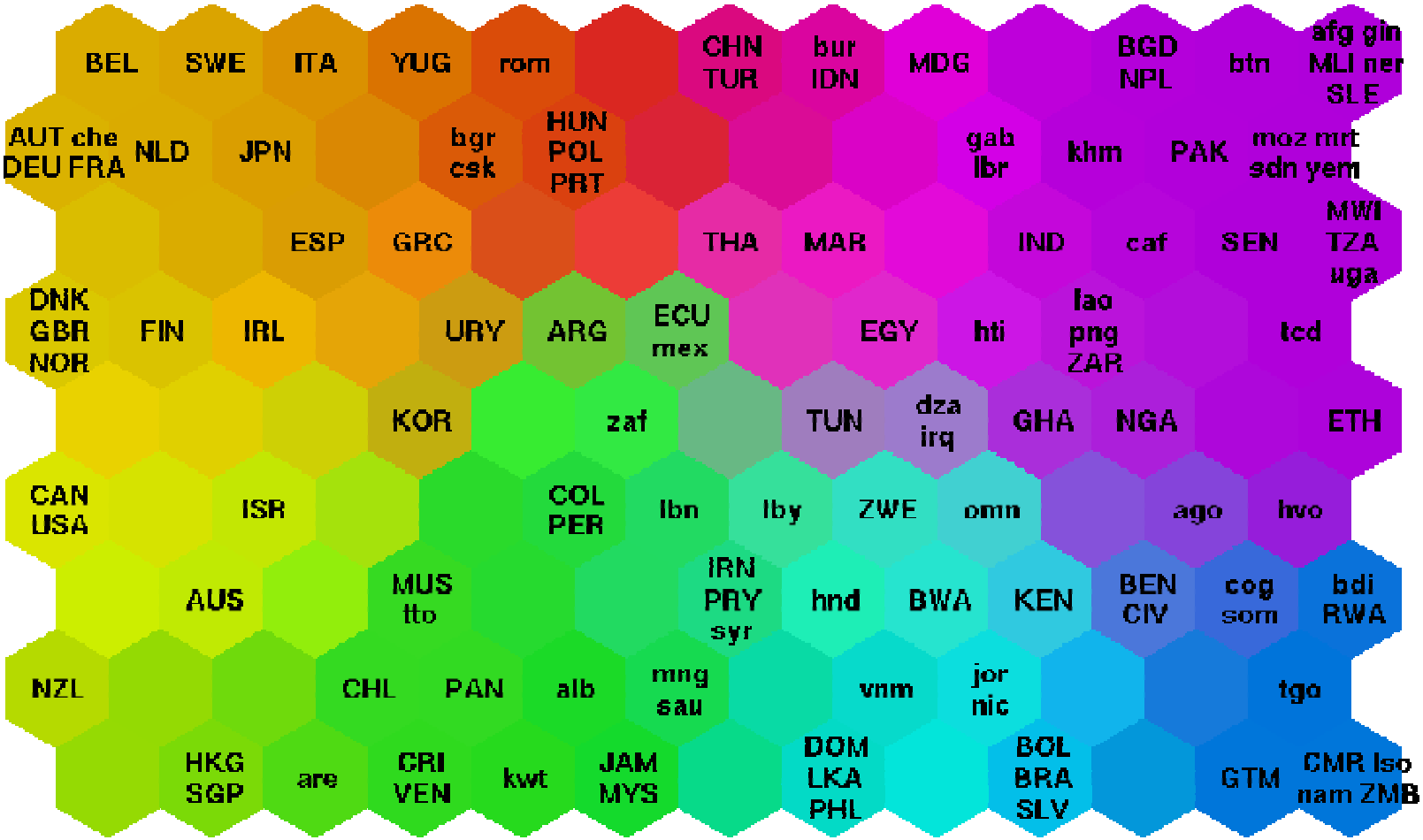
World Bank Statistics

- Data: World Bank statistics of countries in 1992.
- 39 indicators considered e.g., health, nutrition, educational services, etc.
- The complex joint effect of these factors can can be visualized by organizing the countries using the self-organizing map.

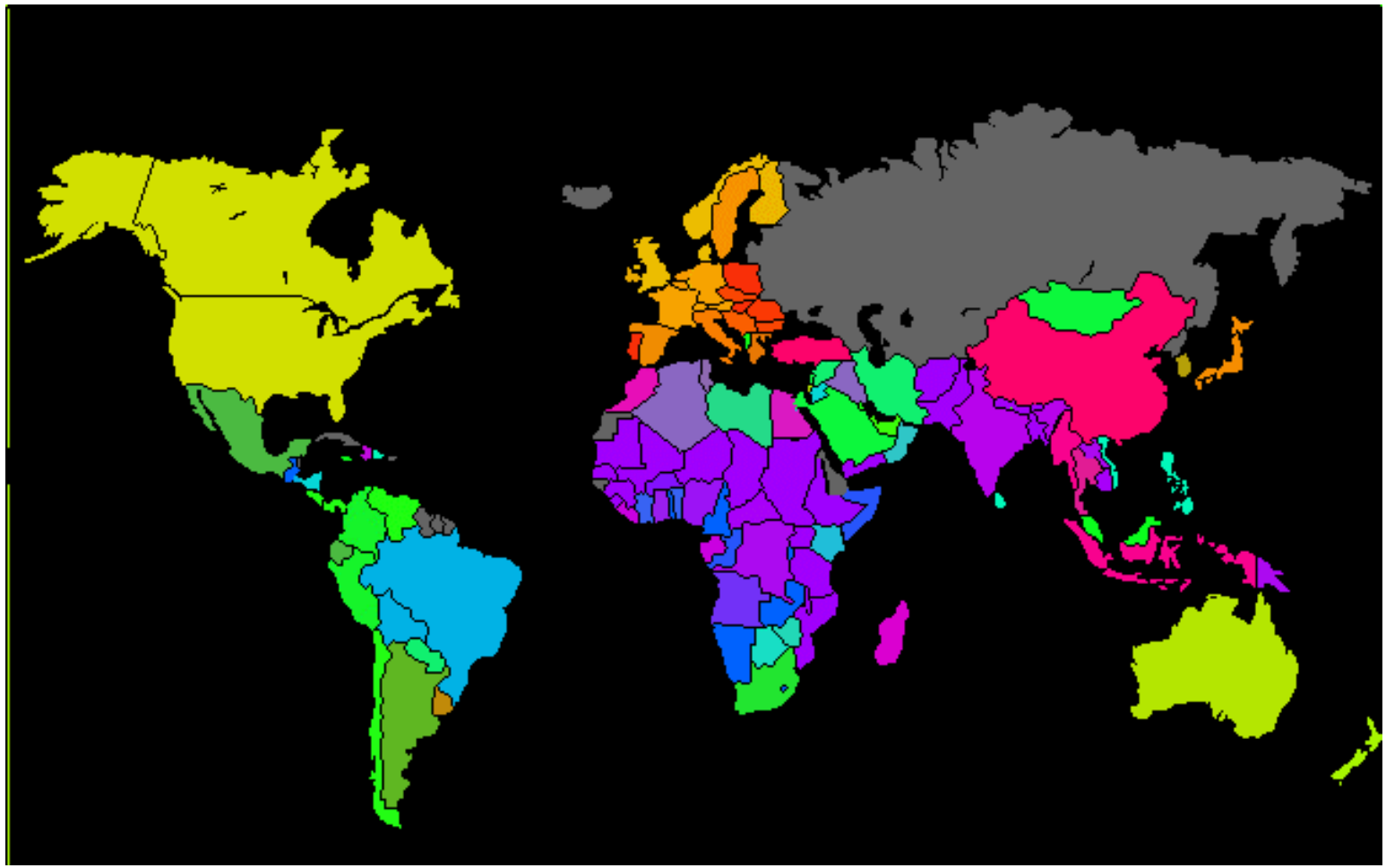
World Poverty PCA

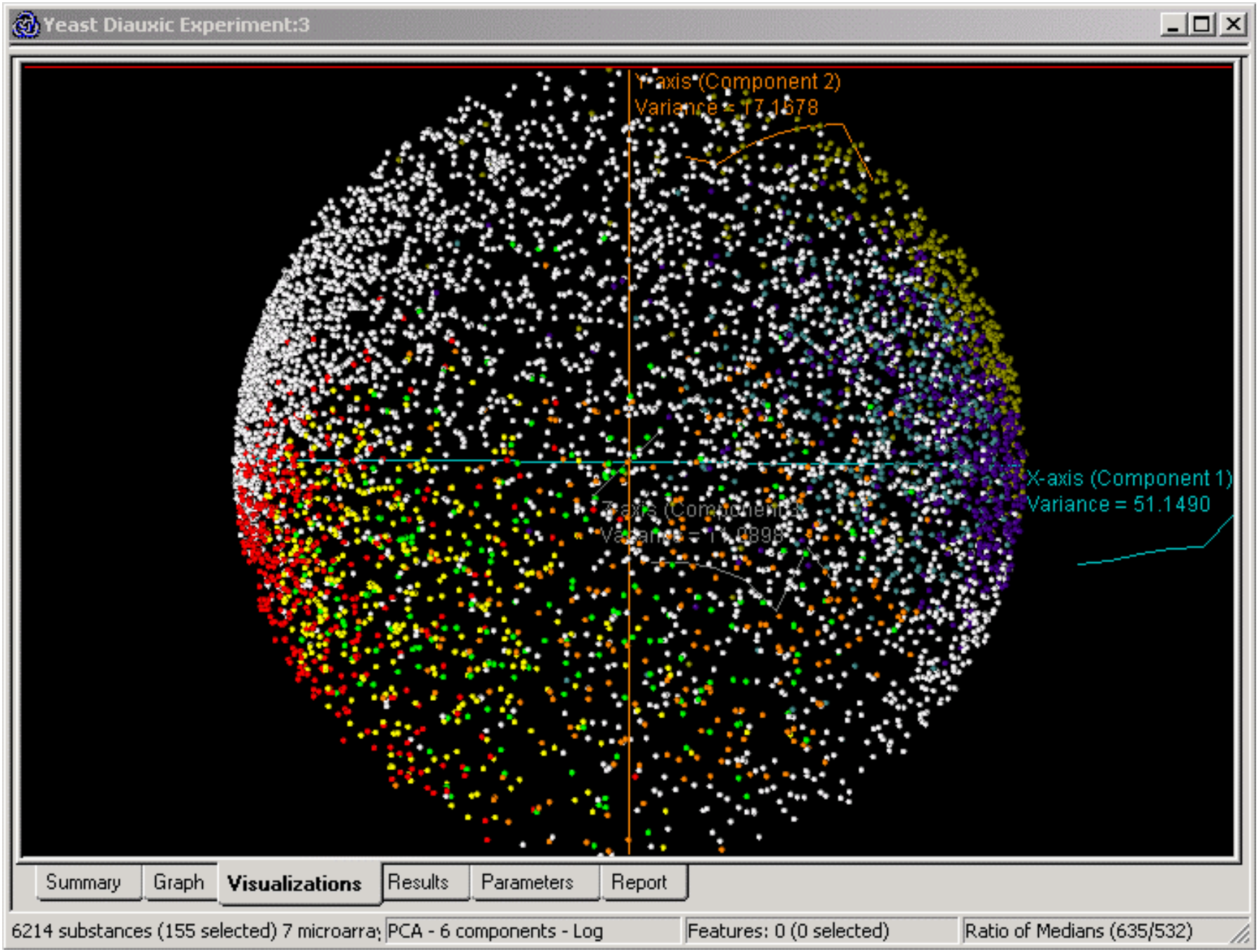


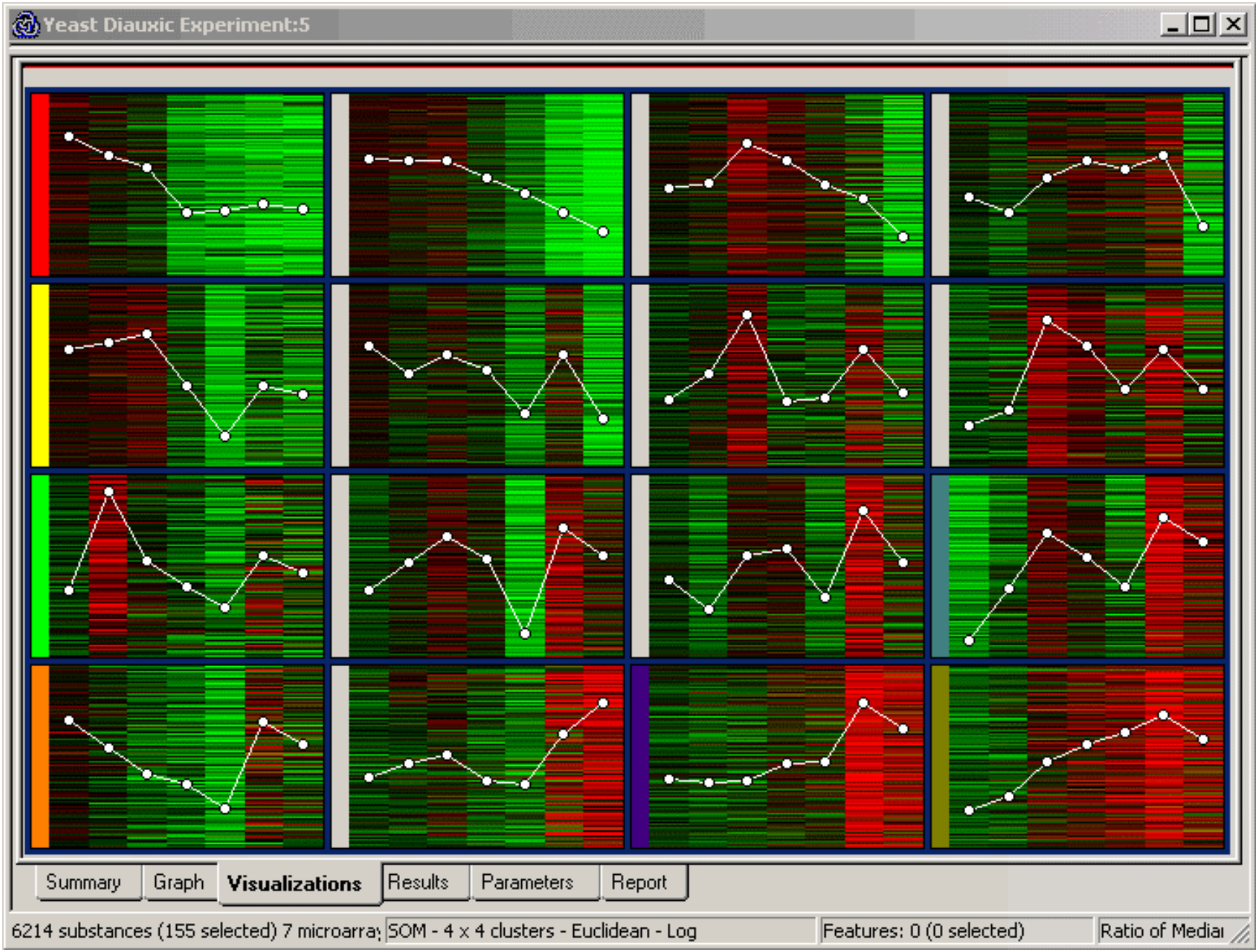
World Poverty SOM



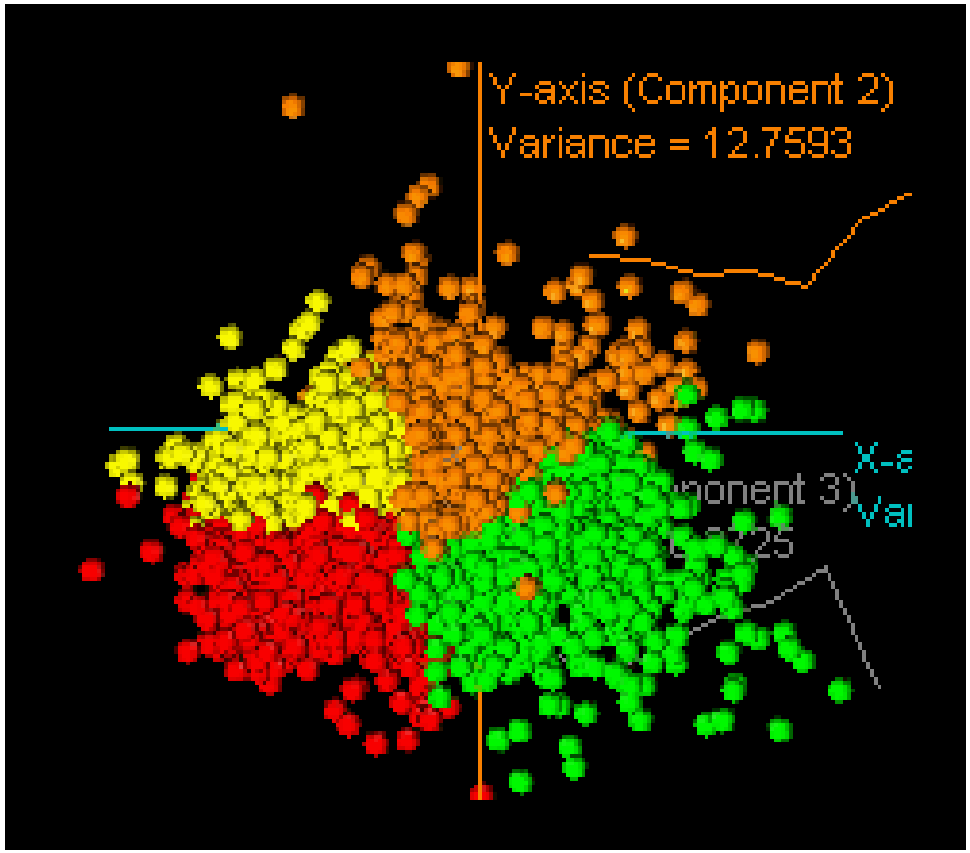
World Poverty Map




Yeast Diauxic Experiment:3
Y-axis (Component 2)
Variance = 17.1678
X-axis (Component 1)
Variance = 51.1490
Z-axis (Component 3)
Variance = 11.0898
Summary
Graph
Visualizations
Results
Parameters
Report
6214 substances (155 selected) 7 microarray
PCA - 6 components - Log
Features: 0 (0 selected)
Ratio of Medians (635/532)


Yeast Diauxic Experiment:5
Summary
Graph
Visualizations
Results
Parameters
Report
6214 substances (155 selected) 7 microarray
SOM - 4 x 4 clusters - Euclidean - Log
Features: 0 (0 selected)
Ratio of Median

Viewing SOM Clusters on PCA axes



SOM Example [Xiao-rui He]

1